

Partial Information Framework: Model-Based Aggregation of Estimates from Diverse Information Sources

Ville A. Satopää, Shane T. Jensen, Robin Pemantle, and Lyle H. Ungar *

Abstract

Prediction polling is an increasingly popular form of crowdsourcing in which multiple participants estimate the probability or magnitude of some future event. These estimates are then aggregated into a single forecast. Historically, randomness in scientific estimation has been generally assumed to arise from unmeasured factors which are viewed as measurement noise. However, when combining subjective estimates, heterogeneity stemming from differences in the participants' information is often more important than measurement noise. This paper formalizes information diversity as an alternative source of such heterogeneity and introduces a novel modeling framework that is particularly well-suited for prediction polls. A practical specification of this framework is proposed and applied to the task of aggregating probability and point estimates from two real-world prediction polls. In both cases our model outperforms standard measurement-error-based aggregators, hence providing evidence in favor of information diversity being the more important source of heterogeneity.

Keywords: Expert belief; Forecast heterogeneity; Judgmental forecasting; Model averaging; Noise reduction

*Ville A. Satopää is a Doctoral Candidate, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6340 (e-mail: satopaa@wharton.upenn.edu); Shane T. Jensen is a Statistician, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6340 (e-mail: stjensen@wharton.upenn.edu); Robin Pemantle is a Mathematician, Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104-6395 (e-mail: pemantle@math.upenn.edu); Lyle H. Ungar is a Computer Scientist, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104-6309 (e-mail: ungar@cis.upenn.edu). This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. The authors would also like to thank Don Moore for providing us with the weight dataset.

1. INTRODUCTION

Past literature has distinguished two types of polling: prediction and opinion polling. In broad terms, an opinion poll is a survey of public opinion, whereas a prediction poll involves multiple agents collectively predicting the value of some quantity of interest (Goel et al., 2010; Mellers et al., 2014). For instance, consider a presidential election poll. An opinion poll typically asks the voters who they will vote for. A prediction poll, on the other hand, could ask which candidate they think will win in their state. A liberal voter in a dominantly conservative state is likely to answer differently to these two questions. Even though opinion polls have been the dominant focus historically, prediction polls have become increasingly popular in the recent years, due to modern social and computer networks that permit the collection of a large number of responses both from human and machine agents. This has given rise to crowdsourcing platforms, such as MTurk and Witkey, and many companies, such as Myriada, Lumenogic, and Inkling, that have managed to successfully capitalize on the benefits of collective wisdom.

This paper introduces statistical methodology designed specifically for the rapidly growing practice of prediction polling. The methods are illustrated on real-world data involving two common types of responses, namely probability and point forecasts. The probability forecasts were collected by the Good Judgment Project (GJP) (Ungar et al. 2012; Mellers et al. 2014) as a means to estimate the likelihoods of international political future events deemed important by the Intelligence Advanced Research Projects Activity (IARPA). Since its initiation in 2011, the project has recruited thousands of forecasters to make probability estimates and update them whenever they felt the likelihoods had changed. To illustrate, Figure 1 shows the forecasts for one of these events. This example involves 522 forecasters making a total of 1,669 predictions between 30 July 2012 and 30 December 2012 when the event finally resolved as “No” (represented by the red line at 0.0). In general, the forecasters reported updates very infrequently. Furthermore, not all forecasters made probability estimates for all the events, making the dataset very sparse. The point forecasts for our second application were collected by Moore

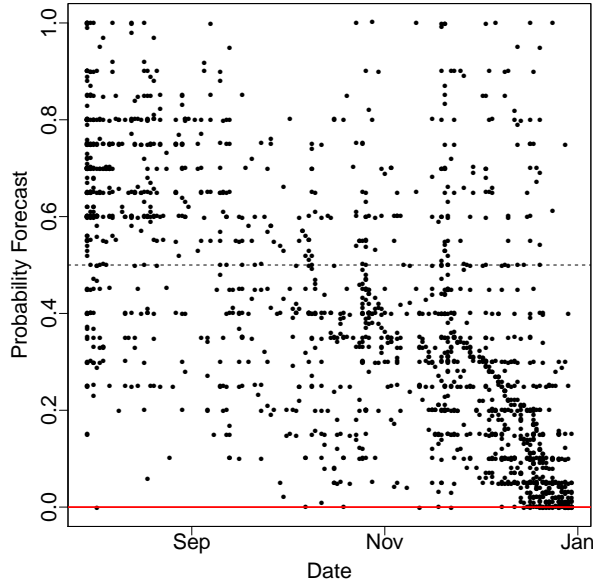


Figure 1: Probability forecasts of the event “Will Moody’s issue a new downgrade on the long-term ratings for any of the eight major French banks between 30 July 2012 and 31 December 2012?” The points have been jittered slightly to make overlaps visible.

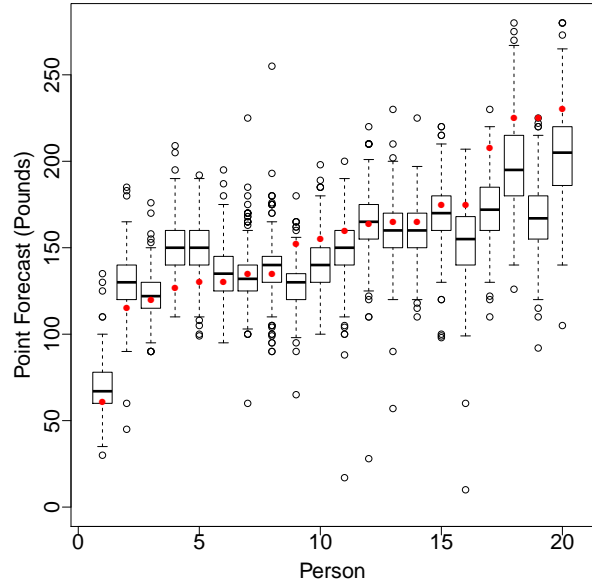


Figure 2: Point forecasts of the weights of 20 different people. The boxplots have been sorted to increase in the true weights (red dots). Some extreme values were omitted for the sake of clarity.

and Klein (2008) who recruited 416 undergraduates from Carnegie Mellon University to guess the weights of 20 people based on a series of pictures. This is an experimental setup where each participant was required to respond to all the questions, leading to a fully completed dataset. The responses are illustrated in Figure 2 that shows the boxplots of the forecasters’ guesses for each of the 20 people. The red dots represent the corresponding true weights.

Once the predictions have been collected, they are typically combined into a single consensus forecast for the sake of decision-making and improved accuracy. Unfortunately, this can be done in many different ways, and the final combination rule can largely determine the out-of-sample performance. The past literature distinguishes two broad approaches to forecast aggregation: empirical aggregation and model-based aggregation. Empirical aggregation is by

far the more widely studied approach; see, e.g., stacking (Breiman, 1996), Bayes model averaging (Raftery et al., 1997), linear opinion pools (DeGroot and Mortera, 1991), and extremizing aggregators (Ranjan and Gneiting, 2010; Satopää et al., 2014a,b). All these methods are akin to machine learning in a sense that they first learn the aggregator based on a training set of past forecasts of known outcomes and then use that aggregator to combine future forecasts of unknown outcomes. Unfortunately, in a prediction polling setup, constructing such a training set requires a lot of effort and time on behalf of the forecasters and the polling agent. Therefore a training set is often not available. Instead, the participants are typically handed a single questionnaire that simultaneously inquires about their predictions of one or more unknown outcomes. This leads to a dataset consisting only of forecasts, which means that empirical aggregation cannot be applied.

Fortunately, model-based aggregation can be performed even when prior knowledge of outcomes is not available. This approach begins by proposing a plausible probability model for the source of heterogeneity among the forecasts, that is, for how and why the forecasts differ from the target outcome. Under this assumed forecast-outcome link, it is then possible to construct an optimal aggregator that can be applied directly to the forecasts without learning the aggregator first from a separate training set. Given this broad applicability, the current paper focuses only on the model-based approach. In particular, outcomes are not assumed available for aggregation at any point in the paper. Instead, aggregation is performed solely based on forecasts, leaving all empirical techniques well outside the scope of the paper.

Historically, potentially due to early forms of data collection, model-based aggregation has considered measurement error as the main source of forecast heterogeneity. This choice motivates aggregators with central tendency such as the (weighted) average, median, and so on. Intuitively, measurement error may be reasonable in modeling repeated estimates from a single instrument. However, it is unlikely to hold in prediction polling, where the estimates arise from multiple, often widely different sources. It is also known that a non-trivial weighted average is

not the optimal aggregator (in terms of the expected quadratic and many other loss functions) under any joint distribution of the outcome and its (conditionally unbiased) forecasts (Dawid et al., 1995; Ranjan and Gneiting, 2010; Satopää and Ungar, 2015). This questions the role of measurement error in model-based aggregation and highlights the need for a different source of forecast heterogeneity.

The main contribution of this paper is a new source of forecast heterogeneity, called *information diversity*, that explains variation by differences in the information available to the forecasters and how they decide to use it. For instance, forecasters studying the same (or different) articles about a company may use separate parts of the information and hence report differing predictions on the company’s future revenue. Such diversity forms the basis of a novel modeling framework known as the *partial information framework*. Theory behind this framework was originally introduced for probability forecasts by Satopää et al. (2015); though their specification is somewhat restrictive for empirical applications. The current paper generalizes the framework beyond probability forecast and removes all unnecessary assumptions, leading to a new specification that is more appropriate for practical applications. This specification allows the decision-maker to build models for different types of forecast-outcome pairs, such as probability forecasts of binary events or point forecasts of real-valued outcomes. Each such model motivates and describes an explicit joint distribution for the target outcome and its forecasts. The optimal aggregator under this joint distribution is available and serves as a more principled model-based alternative to the usual (weighted) average or median.

The paper is structured as follows. Section 2 first describes the partial information framework at its most general level and then introduces a practical specification of the framework. The section ends with a brief review of previous work on model-based aggregation. Section 3 derives a general procedure that guides efficient estimation of the information structure among the forecasters. Section 4 illustrates on real-world data how specific models within the framework can be constructed and applied. In particular, the models are derived and evaluated on

probability and point forecasts from the two prediction polls discussed above. Overall, the resulting partial information aggregators achieve a noticeable performance improvement over the common measurement-error-based aggregators, suggesting that information diversity is the more appropriate model of forecast heterogeneity. Finally, Section 5 concludes with a summary and discussion of future research.

2. MODEL-BASED AGGREGATION

2.1 Bias and Noise

Consider N forecasters and suppose forecaster j predicts X_j for some quantity of interest Y . For instance, in our weight estimation example Y is the true weight of a person and X_j is the guess given by the j th undergraduate. In our probability forecasting application, on the other hand, Y is binary, reflecting whether the event happens or not, and $X_j \in [0, 1]$ is a probability forecast for its occurrence. This section, however, avoids such application specific choices and treats Y and X_j as generic random variables. In general, prediction X_j is nothing but an estimator of Y . Therefore, as is the case with all estimators, its deviation from the truth can be broken down into two components: bias and noise. On the theoretical level, these two components can be separated and hence are often addressed by different mechanisms. This suggests a two-step approach to forecast aggregation: i) eliminate any bias in the forecasts, and ii) combine the unbiased forecasts.

Historically, bias in human judgment has been extensively studied in the psychology literature (for reviews, see Lichtenstein et al. 1977; Yates 1990; Keren 1991). This bias often exhibits well-known patterns (see, e.g., the easy-hard effect in Lichtenstein and Fischhoff 1977; Juslin 1993), and many authors have proposed both cognitive and motivational models to explain it (Koriat et al., 1980; Kruglanski, 1990; Soll, 1996; Moore and Healy, 2008). These models and other results in this popular area of research suggest ways for ex-ante bias reduction. Such

techniques, however, are not in the scope of this paper. Instead, the focus here is on noise reduction and hence specifically on developing methodology for the second step in the overall process of forecast aggregation. In particular, Section 2.2 describes our new framework for modeling the noise component. This is then compared in Section 2.3 to previous noise models. These models make different assumptions about the way the unbiased forecasts relate to the target outcome and hence motivate very different classes of model-based aggregators.

2.2 Partial Information Framework

2.2.1 General Framework

The partial information framework assumes that Y and X_j are measurable under some common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The probability measure \mathbb{P} provides a non-informative yet proper prior on Y and reflects the *basic information* known to all forecasters. Such a prior has been discussed extensively in the economics and game theory literature where it is usually known as the *common prior*. Even though this is a substantive assumption in the framework, specifying a prior distribution cannot be avoided as long as the model depends on a probability space. This includes essentially any probability model for forecast aggregation. How the prior is incorporated depends on the problem context: it can be chosen explicitly by the decision-maker, computed based on past observations of Y , or estimated directly from the forecasts.

The principal σ -field \mathcal{F} can be interpreted as all the possible information that can be known about Y . On top of the basic information reflected in the prior, the j th forecaster uses some personal partial information set $\mathcal{F}_j \subseteq \mathcal{F}$ and predicts $X_j = \mathbb{E}(Y | \mathcal{F}_j)$. Therefore $\mathcal{F}_i \neq \mathcal{F}_j$ if $X_i \neq X_j$, and forecast heterogeneity stems purely from *information diversity*. Note, however, that if forecaster j uses a simple rule, \mathcal{F}_j may not be the full σ -field of information available to the forecaster but rather a smaller σ -field corresponding to the information used by the rule. Furthermore, if two forecasters have access to the same σ -field, they may decide to use different sub- σ -fields, leading to different predictions. This is particularly salient in our weight

estimation example where each forecaster has access to the exact same information, namely the picture of the person, but can choose to use different subsets of this information. Therefore, information diversity does not only arise from differences in the available information, but also from how the forecasters decide to use it. This general point of view was motivated in Satopää et al. (2015) with simple examples that illustrate how the optimal aggregate is not well-defined without assumptions on the information structure among the forecasters.

Satopää et al. (2015) also show that $X_j = \mathbb{E}(Y | \mathcal{F}_j)$ is precisely the same as having a calibrated (sometimes also known as reliable) forecast, that is, $X_j = \mathbb{E}(Y | X_j)$. Therefore the form $X_j = \mathbb{E}(Y | \mathcal{F}_j)$ arises directly from the existence of an underlying probability model and calibration. Overall, calibration $X_j = \mathbb{E}(Y | X_j)$ has been widely discussed in the statistical and meteorological forecasting literature (see, e.g., Dawid et al. 1995; Ranjan and Gneiting 2010; Broecker 2012), with traces at least as far back as Murphy and Winkler (1987). Given that the condition $X_j = \mathbb{E}(Y | X_j)$ depends on the probability measure \mathbb{P} , it should be referred to as \mathbb{P} -calibration when the choice of the probability measure needs to be emphasized. This dependency shows the main conceptual difference between \mathbb{P} -calibration and the notion of empirical calibration (Dawid 1982; Foster and Vohra 1998; and many others). However, as was pointed out by Dawid et al. (1995), these two notions can be expressed in formally identical terms by letting \mathbb{P} represent the limiting joint distribution of the forecast-outcome pairs.

In practice researchers have discovered many calibrated subpopulations of experts, such as meteorologists (Murphy and Winkler, 1977a,b), experienced tournament bridge players (Keren, 1987), and bookmakers (Dowie, 1976). Generally, calibration can be improved through team collaboration, training, tracking (Mellers et al., 2014), performance feedback (Murphy and Daan, 1984), representative sampling of target events (Gigerenzer et al., 1991; Juslin, 1993), or by evaluating the forecasters' performance under a loss function that is minimized by the conditional expectation of Y , given the forecaster's information (Banerjee et al., 2005). If one is nonetheless left with uncalibrated forecasts, they can be calibrated ex-ante as follows.

First, consider some (possibly uncalibrated) forecasts $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_N)'$ defined on (Ω, \mathcal{F}) . Choose some distribution \mathbb{Q} for $(Y, \tilde{\mathbf{X}})$. For instance, Dawid et al. 1995 suggest first choosing a distribution \mathbb{Q} for $\tilde{\mathbf{X}}$ and then setting $\mathbb{Q}(Y, \tilde{\mathbf{X}}) = \Psi(\tilde{\mathbf{X}})\mathbb{Q}(\tilde{\mathbf{X}})$, where Ψ is an arbitrary aggregator (such as the average of probability forecasts of a binary event) acting as $\mathbb{Q}(Y|\tilde{\mathbf{X}})$. Alternatively, one may search for an appropriate \mathbb{Q} in the large literature of quantitative psychology. Regardless how \mathbb{Q} is constructed, however, the calibrated version of \tilde{X}_j is $\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j)$. This forecast is \mathbb{Q} -calibrated and can be written as $\mathbb{E}_{\mathbb{Q}}(Y|\mathcal{F}_j)$, where $\mathcal{F}_j = \sigma(\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j))$ is the σ -field generated by $\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j)$. Intuitively, calibrating is equivalent to replacing forecast x by $\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j = x)$ for all possible values $x \in \text{supp}(\tilde{X}_j)$. Perhaps, however, one does not want to work under this particular model. To accommodate alternative models (such as the Gaussian model described in Section 2.2.2), the next proposition shows how \mathbb{Q} -calibrated forecasts can be transformed into forecasts that are calibrated under some other probability measure \mathbb{P} . All the proofs are deferred to Appendix A.

Proposition 2.1. *Consider a probability measure \mathbb{P} such that $\mathbb{P} \ll \mathbb{Q}$. Let $\frac{d\mathbb{P}}{d\mathbb{Q}}$ denote the Radon-Nikodym derivative of \mathbb{P} with respect to \mathbb{Q} . The forecasts under the new model \mathbb{P} are then given by the transformation $\mathbb{E}_{\mathbb{P}}(Y|\mathcal{F}_j) = \mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}Y|\mathcal{F}_j\right) / \mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}|\mathcal{F}_j\right)$, where $\mathcal{F}_j = \sigma(\mathbb{E}_{\mathbb{Q}}(Y|\tilde{X}_j))$.*

This shows that uncalibrated forecasts from “non-experts” can be calibrated as long as one agrees on some joint distribution for the target outcome and its forecasts. While such constructs certainly deserve further analysis, they are not in the scope of this paper and hence are left for future work. Therefore, from now on, the forecasts are assumed to be calibrated. Note, however, that in general the forecasts should satisfy some minimal performance criterion; simply aggregating entirely arbitrary forecasts is hardly going to lead to improved forecasting accuracy. To this end, Foster and Vohra (1998) analyze probability forecasts and state that “calibration does seem to be an appealing minimal property that any probability forecast should satisfy.” They show that one needs to know almost nothing about the outcomes in order to be calibrated. Thus, in theory, calibration can be achieved very easily and overall seems like an

appropriate base assumption for developing a general theory of forecast aggregation.

Given that the partial information framework generates all forecast variation from information diversity, it is important to understand the extent to which the forecasters' partial information sets can be measured in practice. First, note that, for the purposes of aggregation, any available information discarded by a forecaster may as well not exist because information comes to the aggregator only through the forecasts. Therefore it is not in any way restrictive to assume that $\mathcal{F}_j = \sigma(X_j)$. Second, the following proposition describes observable measures for the amount of information in each forecast and for the amount of information overlap between any two forecasts.

Proposition 2.2. *If $\mathcal{F}_j = \sigma(X_j)$ such that $\mathbb{E}(Y|\mathcal{F}_j) = \mathbb{E}(Y|X_j) = X_j$ for all $j = 1, \dots, N$, then the following holds.*

- i) Forecasts are marginally consistent: $\mathbb{E}(Y) = \mathbb{E}(X_j)$.*
- ii) Variance increases in information: $\text{Var}(X_i) \leq \text{Var}(X_j)$ if $\mathcal{F}_i \subseteq \mathcal{F}_j$. Given that $Y = \mathbb{E}(Y|\mathcal{F})$, the variances of the forecasts are upper bounded as $\text{Var}(X_j) \leq \text{Var}(Y)$ for all $j = 1, \dots, N$.*
- iii) $\text{Cov}(X_j, X_i) = \text{Var}(X_i)$ if $\mathcal{F}_i \subseteq \mathcal{F}_j$. Again, expressing $Y = \mathbb{E}(Y|\mathcal{F})$ implies that $\text{Cov}(X_j, Y) = \text{Var}(X_j)$ for all $j = 1, \dots, N$.*

This proposition is important for multiple reasons. First, item i) provides guidance in estimating the prior mean of Y from the observed forecasts. Second, item ii) shows that $\text{Var}(X_j)$ quantifies the amount of information used by forecaster j . In particular, $\text{Var}(X_j)$ increases to $\text{Var}(Y)$ as forecaster j learns and becomes more informed. Therefore increased variance reflects more information and is deemed helpful. This is a clear contrast to the standard statistical models that often regard higher variance as increased noise and hence harmful. The covariance $\text{Cov}(X_i, X_j)$, on the other hand, can be interpreted as the amount of information overlap between forecasters i and j . Given that being non-negatively correlated is not generally transitive

(Langford et al., 2001), these covariances are not necessarily non-negative even though all forecasts are non-negatively correlated with the outcome. Such negatively correlated forecasts can arise in a real-world setting. For instance, consider two forecasters who see voting preferences of two different sub-populations that are politically opposed to each other. Each individually is a weak predictor of the total vote on any given issue, but they are negatively correlated because of the likelihood that these two blocks will largely oppose each other.

Third and finally, item iii) shows that the covariance matrix Σ_X of the X_j s extends to the unknown Y as follows:

$$\text{Cov}((Y, X_1, \dots, X_N)') = \begin{pmatrix} \text{Var}(Y) & \text{diag}(\Sigma_X)' \\ \text{diag}(\Sigma_X) & \Sigma_X \end{pmatrix}, \quad (1)$$

where $\text{diag}(\Sigma_X)$ denotes the diagonal of Σ_X . This is the key to regressing Y on the X_j s without a separate training set of past forecasts of known outcomes. The resulting estimator, called the *revealed aggregator*, is

$$X'' := \mathbb{E}(Y|X_1, \dots, X_N) = \mathbb{E}(Y | \mathcal{F}''),$$

where $\mathcal{F}'' := \sigma(X_1, \dots, X_N)$ is the σ -field generated (or information revealed) by the X_j s. The revealed aggregator uses all the information that is available in the forecasts and hence is the optimal aggregator under the distribution of (Y, X_1, \dots, X_N) . To make this precise, consider a scoring rule $S(x, y)$ that represents the loss of predicting x when the outcome is y . A scoring rule is said to be consistent for the mean of Y if $\mathbb{E}_Y[S(\mathbb{E}_Y(Y), Y)] \leq \mathbb{E}_Y[S(x, Y)]$ for all $x \in \mathbb{R}$. Savage (1971) showed, subject to weak regularity conditions, that all such scoring rules can be written in the form

$$S(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x), \quad (2)$$

where ϕ is a convex function with subgradient ϕ' . An important special case is the quadratic loss $S(x, y) = (x - y)^2$ that arises when $\phi(x) = x^2$. Now, if an aggregator is defined as any random variable $X \in \sigma(X_1, \dots, X_N)$, then X'' is an aggregator that minimizes expectation of any scoring rule S of the form (2):

$$\begin{aligned}\mathbb{E}[S(X, Y)] &= \mathbb{E}_{X_1, \dots, X_N} \{ \mathbb{E}_{Y|X_1, \dots, X_N} [S(X, Y)] \} \\ &\geq \mathbb{E}_{X_1, \dots, X_N} \{ \mathbb{E}_{Y|X_1, \dots, X_N} [S(X'', Y)] \} \\ &= \mathbb{E}[S(X'', Y)].\end{aligned}$$

Ranjan and Gneiting (2010) showed a similar results for probability forecasts. For these reasons, X'' is considered the relevant aggregator under each specific instance of the framework. The next section shows how this aggregator can be captured in practice.

2.2.2 Gaussian Partial Information Model

Even though the general framework is convenient for theoretical analysis, it is clearly too abstract for practical applications. Fortunately, applying the framework in practice only requires one extra assumption, namely the choice of a parametric family for the distribution of (Y, X_1, \dots, X_N) . One approach is to refer to Proposition 2.2 and choose a family that is parametrized in terms of the first two joint moments. This points at the multivariate Gaussian distribution that is a typical starting point in developing statistical methodology and often provides the cleanest entry into the issues at hand.

The Gaussian distribution is also the most common choice for modeling measurement error. This is typically motivated by assuming the terms to represent sums of a large number of independent sources of error. The central limit theorem then gives a natural motivation for the Gaussian distribution. A similar argument can be made under the partial information framework. First, consider some pieces of information. Each piece either has a positive or negative

impact and hence respectively either increases or decreases Y . The total sum (integral) of these pieces determines the value of Y . Each forecaster, however, only observes the sum of some subset of them. Based on this sum, the forecaster makes an estimate of Y . If the pieces are independent and have small tails, then the joint distribution of the forecasters' observations will be asymptotically Gaussian. Given that the number of information pieces in a real-world setup is likely to be large, it makes sense to model the forecasters' observations as jointly Gaussian. Of course, other distributions, such as the multivariate t -distribution, are possible. At this point, however, such alternative specifications are best left for future work.

The model variables (Y, X_1, \dots, X_N) can be modeled directly with a Gaussian distribution as long as they are all real-valued. In many applications, however, Y and X_j may not be supported on the whole real line. For instance, the aforementioned Good Judgment Project collected probability forecasts of binary events. In this case, $X_j \in [0, 1]$ and $Y \in \{0, 1\}$. Fortunately, different types of outcome-forecast pairs can be easily addressed by borrowing from the theory of generalized linear models (McCullagh and Nelder, 1989) and utilizing a *link function*. The result is a close yet widely applicable specification called the *Gaussian partial information model*. This model begins by introducing $N + 1$ information variables that follow a multivariate Gaussian distribution with the covariance pattern (1):

$$\begin{pmatrix} Z_0 \\ Z_1 \\ \vdots \\ Z_N \end{pmatrix} \sim \mathcal{N}_{N+1} \left(\mathbf{0}, \begin{pmatrix} 1 & \text{diag}(\boldsymbol{\Sigma})' \\ \text{diag}(\boldsymbol{\Sigma}) & \boldsymbol{\Sigma} \end{pmatrix} := \left(\begin{array}{c|cccc} 1 & \delta_1 & \delta_2 & \dots & \delta_N \\ \hline \delta_1 & \delta_1 & \rho_{1,2} & \dots & \rho_{1,N} \\ \delta_2 & \rho_{2,1} & \delta_2 & \dots & \rho_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_N & \rho_{N,1} & \rho_{N,2} & \dots & \delta_N \end{array} \right) \right). \quad (3)$$

This distribution supports the Gaussian model similarly to the way the ordinary linear regression supports the class of generalized linear models. In particular, the information variables transform into the outcome and forecasts via an application-specific link function $g(\cdot)$; that

is, $Y = g(Z_0)$ and $X_j = \mathbb{E}(Y|Z_j) = \mathbb{E}(g(Z_0)|Z_j)$. Given that Z_0 fully determines Y , it is sufficient for all information that can be known about Y . The remaining variables Z_1, \dots, Z_N , on the other hand, summarize the forecasters' partial information. To make this more concrete, consider our two real-world applications. For probability forecasts of a binary event a reasonable link function $g(\cdot)$ is the indicator function $\mathbf{1}_A$, where $A = \{Z_0 > t\}$ for some threshold value $t \in \mathbb{R}$. For real-valued X_j and Y , on the other hand, a reasonable choice is the reverse standardizing function $g(Z_0) = \sigma_0 Z_0 + \mu_0$, where μ_0 and σ_0 are the prior mean and standard deviation of Y , respectively. In general, it makes sense to have $g(\cdot)$ map from the real-numbers to the support of Y such that Y has the correct prior $\mathbb{P}(Y)$.

Overall, this model can be considered as a close yet practical specification of the general framework. After all, it only adds on the assumption of Gaussianity. This extra assumption, however, is enough to allow the construction of the revealed aggregator $X'' = \mathbb{E}(Y|Z_1, \dots, Z_N)$. For X'' and also X_j the conditional expectations can be often computed via the following conditional distributions:

$$\begin{aligned} Z_0|Z_j &\sim \mathcal{N}(Z_j, 1 - \delta_j) \text{ and} \\ Z_0|\mathbf{Z} &\sim \mathcal{N}(\text{diag}(\Sigma)' \Sigma^{-1} \mathbf{Z}, 1 - \text{diag}(\Sigma)' \Sigma^{-1} \text{diag}(\Sigma)), \end{aligned}$$

where $\mathbf{Z} = (Z_1, \dots, Z_N)'$. For instance, if both X_j and Y are real-valued, then $X_j = \sigma_0 Z_j + \mu_0$ and $X'' = \text{diag}(\Sigma)' \Sigma^{-1} (\mathbf{X} - \mu_0 \mathbf{1}_N) + \mu_0$, where $\mathbf{X} = (X_1, \dots, X_N)'$. These conditional distributions arise directly from the well-known conditional distributions of the multivariate Gaussian distribution (see, e.g., Ravishanker and Dey 2001).

2.3 Previous Work on Model-Based Aggregation

2.3.1 Interpreted Signal Framework

The *interpreted signal framework* is a behavioral model that assumes different predictions to arise from differing interpretation procedures (Hong and Page, 2009). For example, consider two forecasters who visit a company and predict its future revenue. One forecaster may carefully examine the company’s technological status while the other pays closer attention to what the managers say. Even though the forecasters receive and possibly even use the exact same information, they may interpret it differently and hence end up reporting different forecasts. Therefore forecast heterogeneity is assumed to stem from “cognitive diversity”.

This is a very reasonable model and hence has been used in various forms to simulate and illustrate theory about expert behavior (see, e.g., Broomell and Budescu 2009; Parunak et al. 2013). Consequently, previous authors have constructed many highly specialized toy models of interpreted forecasts. For instance, Dawid et al. (1995) construct simple models of two forecasts to support their discussion on coherent forecast aggregation; Ranjan and Gneiting (2010) use one of these models to simulate calibrated forecasts; and Di Bacco et al. (2003) introduce a model for two forecasters whose (interpreted) log-odds predictions follow a joint Gaussian distribution. Unfortunately, their model is very narrow due to its detailed assumptions and extensive computations. Furthermore, it is not clear how the model can be used in practice or extended to N forecasters. All in all, it seems that successful previous applications of the interpreted signal framework have used it as a basis for illustrating theory instead of actually aiming to model real-world forecasts. In this respect, the framework has remained relatively abstract.

Our partial information framework, however, formalizes the intuition behind it, allows quantitative predictions, and provides a flexible construction for modeling many different forecasting setups. Overall, the framework is very general and, in fact, encompasses all the other

authors’ models mentioned above as different sub-cases. Unlike the Gaussian model, however, these models make many restrictive assumptions in addition to just choosing a parametric family. Even though the general partial information framework, as described in Section 2.2, does not allow the forecasters to interpret information differently and hence does not capture all aspects of the interpreted signal framework, personal interpretations can be easily introduced by associating forecaster j with a probability measure \mathbb{P}_j that describes that forecaster’s interpretation of information. If \mathbb{E}_j denotes the expectation under \mathbb{P}_j , then it is possible that $X_i = \mathbb{E}_i(Y|\mathcal{F}_i) \neq X_j = \mathbb{E}_j(Y|\mathcal{F}_j)$ even if $\mathcal{F}_i = \mathcal{F}_j$. In practice, however, eliciting the details of each \mathbb{P}_j is hardly possible. Therefore, to keep the model tractable, it is convenient to assume a common interpretation $\mathbb{P}_j = \mathbb{P}$ for all $j = 1, \dots, N$.

2.3.2 Measurement Error Framework

In the absence of a quantitative interpreted signal model, prior applications have typically explained forecast heterogeneity with standard statistical models. These models are different formalizations of the *measurement error framework* that generates forecast heterogeneity purely from a probability distribution. More specifically, this framework assumes a “true” (possibly transformed) forecast θ , which can be interpreted as the prediction made by an ideal forecaster. The forecasters then somehow measure θ with mean-zero idiosyncratic error. For instance, in our probability forecasting application one possible measurement error model is

$$\begin{aligned} Y &\sim \text{Bernoulli}(\theta), \\ \text{logit}(X_j) &= \text{logit}(\theta) + e_j, \text{ and} \\ e_j &\stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \text{ for all } j = 1, \dots, N, \end{aligned} \tag{4}$$

where $\text{logit}(x) = \log(x/(1-x))$ is the log-odds operator. Given that the errors are generally assumed to have mean zero, measurement error forecasts are unbiased estimates of θ , that is,

$\mathbb{E}(X_j|\theta) = \theta$. Observe that this is not the same as assuming calibration $\mathbb{E}(Y|X_j) = X_j$. Therefore an unbiased estimation model is very different from a calibrated model. This distinction is further emphasized by the fact that X'' never reduces to a (non-trivial) weighted average of the forecasts (Satopää and Ungar, 2015). Given that the measurement-error aggregators are often different types of weighted averages, measurement error and information diversity are not only philosophically different but they also require very different aggregators.

Example (4) illustrates the main advantages of the measurement error framework: simplicity and familiarity. Unfortunately, there are a number of disadvantages. First, measurement-error aggregators estimate θ instead of the realized value of the random variable Y . For this reason, these aggregators often do not satisfy even the minimal performance requirements. For instance, a non-trivial weighted average of calibrated forecasts is necessarily both uncalibrated and under-confident (Ranjan and Gneiting, 2010; Satopää and Ungar, 2015). Second, the standard assumption of conditional independence of the observations forces a specific and highly unrealistic structure on interpreted forecasts (Hong and Page, 2009). Measurement-error aggregators also cannot leave the convex hull of the individual forecasts, which further contradicts the interpreted signal framework (Parunak et al., 2013) and can be easily seen to result in poor empirical performance on many datasets. Third, the underlying model is rather implausible. Relying on a true forecast θ invites philosophical debate, and even if one assumes the existence of such a value, it is difficult to believe that the forecasters are actually seeing it with independent noise. Therefore, whereas the interpreted signal framework proposes a plausible micro-level explanation, the measurement error model does not; at best, it forces us to imagine a group of forecasters who apply the same procedures to the same data but with numerous small mistakes.

3. MODEL ESTIMATION

This section describes methodology for estimating the *information structure* Σ . Even though Σ is mostly used for aggregation, it also describes the information among the forecasters (see end of Section 2.2.1) and hence should be of interest to decision analysts, psychologists, and the broader community studying collective problem solving. Unfortunately, estimating Σ in full generality based on a single prediction per forecaster is difficult. Therefore, to facilitate model estimation, the forecasters are assumed to predict $K \geq 2$ related events. For instance, in our second application 416 undergraduates guessed the weights of 20 people. This yielded a 20×416 matrix that was then used to estimate Σ .

3.1 General Estimation Problem

Denote the outcome of the k th event with Y_k and the j th forecaster's prediction for this outcome with X_{jk} . For the sake of generality, this section does not assume any particular link function but instead operates directly with the corresponding information variables, denoted with Z_{jk} . In practice, the forecasts X_{jk} can be often transformed into Z_{jk} at least approximately. This is illustrated in Section 4. Recall that aggregation cannot access to the outcomes $\{Y_1, \dots, Y_K\}$ or their corresponding information variables $\{Z_{01}, \dots, Z_{0K}\}$. Instead, Σ is estimated only based on $\{\mathbf{Z}_1, \dots, \mathbf{Z}_K\}$, where the vector $\mathbf{Z}_k = (Z_{1k}, \dots, Z_{Nk})'$ collects the forecasters' information about the k th event.

This estimation must respect the covariance pattern (3). More specifically, if \mathcal{S}_+^N denotes the set of $N \times N$ symmetric positive semidefinite matrices and

$$h(\mathbf{M}) := \begin{pmatrix} 1 & \text{diag}(\mathbf{M})' \\ \text{diag}(\mathbf{M}) & \mathbf{M} \end{pmatrix}$$

for some symmetric matrix \mathbf{M} , then the final estimate must satisfy the condition $h(\Sigma) \in \mathcal{S}_+^{N+1}$.

Intuitively, this is satisfied if there exists a random variable Y for which the forecasts X_j are jointly calibrated. In terms of information, this means that it is physically possible to allocate information about Y among the N forecasters in the manner described by Σ . Therefore the condition is named *information coherence*.

Unfortunately, simply finding an accurate estimate of Σ does not guarantee precise aggregation. To see this, recall from Section 2.2.2 that $\mathbb{E}(Z_{0k}|\mathbf{Z}_k) = \text{diag}(\Sigma)' \Sigma^{-1} \mathbf{Z}_k$. This term is generally found in the revealed aggregator and hence deserves careful treatment. Re-express the term as $\mathbf{v}' \mathbf{Z}_k$, where \mathbf{v} is the solution to $\text{diag}(\Sigma) = \Sigma \mathbf{v}$. The rate at which the solution changes with respect to a change in $\text{diag}(\Sigma)$ depends on the condition number $\text{cond}(\Sigma) := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, i.e., the ratio between the maximum and minimum eigenvalues of Σ . If the condition number is very large, a small error in $\text{diag}(\Sigma)$ can cause a large error in \mathbf{v} . If the condition number is small, Σ is called *well-conditioned* and error in \mathbf{v} will not be much larger than the error in $\text{diag}(\Sigma)$. Thus, to prevent estimation error from being amplified during aggregation, the estimation procedure should require $\text{cond}(\Sigma) \leq \kappa$ for a given threshold $\kappa \geq 1$.

This all gives the following general estimation problem:

$$\begin{aligned} & \text{minimize } f_0(\Sigma, \{\mathbf{Z}_1, \dots, \mathbf{Z}_k\}) \\ & \text{subject to } h(\Sigma) \in \mathcal{S}_+^{N+1}, \text{ and} \\ & \text{cond}(\Sigma) \leq \kappa, \end{aligned} \tag{5}$$

where f_0 is some objective function. The feasible region defined by the two constraints is convex. Therefore, if f_0 is convex in Σ , expression (5) is a convex optimization problem. Typically the global optimum to such a problem can be found very efficiently. Problem (5), however, involves $\binom{N+1}{2}$ variables. Therefore it can be solved efficiently with standard optimization techniques, such as the interior point methods, as long as the number of variables is not too large, say, not more than 1,000. Unfortunately, this means that the procedure cannot be

applied to prediction polls with more than about $N = 45$ forecasters. This is very limiting as many prediction polls involve hundreds of forecasters. For instance, our two real-world applications involve 100 and 416 forecasters. Fortunately, by choosing the loss function carefully one can perform dimension reduction and estimate Σ under a much larger N . This is illustrated in the following subsections.

3.2 Maximum Likelihood Estimator

Under the Gaussian model the information structure Σ is a parameter of an explicit likelihood. Therefore estimation naturally begins with the maximum likelihood approach (MLE). Unfortunately, the Gaussian likelihood is not convex in Σ . Consequently, only a locally optimal solution is guaranteed with standard optimization techniques. Furthermore, it is not clear whether the dimension of this form can be reduced. Won and Kim (2006) discuss the MLE under a condition number constraint. They are able to transform the original problem with $\binom{N+1}{2}$ variables to an equivalent problem with only N variables, namely the eigenvalues of Σ . This transformation, however, requires an orthogonally invariant problem. Given that the constraint $h(\Sigma) \in \mathcal{S}_+^{N+1}$ is not orthogonally invariant, the same dimension-reduction technique cannot be applied. Instead, the MLE must be computed with the $\binom{N+1}{2}$ variables, making estimation slow for small N and undoable even for moderately large N . For these reasons the MLE is not discussed further in this paper.

3.3 Least Squares Estimator

Past literature has discussed many simple covariance estimators that can be applied efficiently to large amounts of data. Unfortunately, these estimators are not guaranteed to satisfy the conditions in (5). This section introduces a correctional procedure that inputs any covariance estimator S and modifies it minimally such that the end result satisfies the conditions in (5). More specifically, S is projected onto the feasible region. This approach, sometimes known as

the least squares approach (LSE), motivates a convex loss function that guarantees a globally optimal solution and facilitates dimension reduction. Most importantly, however, it provides a general tool for estimating Σ , regardless whether one is working with a Gaussian model or possibly some future non-Gaussian model.

From the computational perspective, it is more convenient to project $h(\mathbf{S})$ instead of \mathbf{S} . Even though this could be done under many different norms, for the sake of simplicity, this paper only considers the squared Frobenius norm $\|\mathbf{M}\|_F^2 = \text{tr}(\mathbf{M}'\mathbf{M})$, where $\text{tr}(\cdot)$ is the trace operator. The LSE is then given by $h^{-1}(\Omega)$, i.e., Ω without the first row and column, where Ω is the solution to

$$\begin{aligned} & \text{minimize } \|\Omega - h(\mathbf{S})\|_F^2 \\ & \text{subject to } \Omega \in \mathcal{S}_+^{N+1}, \\ & \quad \text{cond}(\Omega) \leq \kappa, \text{ and} \\ & \quad \text{tr}(\mathbf{A}_j \Omega) = b_j, \quad (j = 1, \dots, N+1). \end{aligned} \tag{6}$$

Both \mathbf{A}_j and b_j are constants defined to maintain the covariance pattern (3). More specifically, if \mathbf{e}_j denotes the j th standard basis vector of length $N+1$, then

$$\begin{aligned} b_1 &= 1, \mathbf{A}_1 = \mathbf{e}_1 \mathbf{e}_1', \text{ and} \\ b_j &= 0, \mathbf{A}_j = \mathbf{e}_j \mathbf{e}_j' - 0.5(\mathbf{e}_1 \mathbf{e}_j' + \mathbf{e}_j \mathbf{e}_1') \text{ for } j = 2, \dots, N+1. \end{aligned}$$

If Ω satisfies the other two conditions, namely $\Omega \in \mathcal{S}_+^{N+1}$ and $\text{cond}(\Omega) \leq \kappa$, then $\Sigma = h^{-1}(\Omega)$ also satisfies them. This follows from the fact that Σ is a principal sub-matrix of Ω . Therefore $\Omega \in \mathcal{S}_+^{N+1}$ implies $\Sigma \in \mathcal{S}_+^N$. Furthermore, Cauchy's interlace theorem (see, e.g., Hwang 2004) states that $\lambda_{\min}(\Omega) \leq \lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma) \leq \lambda_{\max}(\Omega)$ such that $\text{cond}(\Sigma) \leq \text{cond}(\Omega) \leq \kappa$. Of course, requiring $\text{cond}(\Omega) \leq \kappa$ instead of $\text{cond}(\Sigma) \leq \kappa$ shrinks the region of feasible Σ s. At this point, however, the exact value of κ is arbitrary and merely serves to control $\text{cond}(\Sigma)$.

Section 3.4 introduces a procedure for choosing κ from the data. Under such an adaptive procedure, problem (6) can be considered equivalent to directly projecting \mathbf{S} onto the feasible region.

The first step towards solving (6) is to express the feasible region as an intersection of the following two sets:

$$\begin{aligned}\mathcal{C}_{sd} &= \{ \mathbf{\Omega} : \mathbf{\Omega} \in \mathcal{S}_+^{N+1}, \text{cond}(\mathbf{\Omega}) \leq \kappa \}, \text{ and} \\ \mathcal{C}_{lin} &= \{ \mathbf{\Omega} : \text{tr}(\mathbf{A}_j \mathbf{\Omega}) = b_j, j = 1, \dots, N+1 \}.\end{aligned}$$

Given that both of these sets are convex, projecting onto their intersection can be computed with the Directional Alternating Projection Algorithm (Gubin et al., 1967). This method makes progress by repeatedly projecting onto the sets \mathcal{C}_{sd} and \mathcal{C}_{lin} . Consequently, it is efficient only if projecting onto each of the individual sets is fast. Fortunately, as will be shown next, this turns out to be the case.

First, projecting an $(N+1) \times (N+1)$ symmetric matrix $\mathbf{M} = \{m_{ij}\}$ onto \mathcal{C}_{lin} is a linear map. To make this more specific, let $\mathbf{m} = \text{vec}(\mathbf{M})$ be a column-wise vectorization of \mathbf{M} . If \mathbf{A} is a matrix with the j th row equal to $\text{vec}(\mathbf{A}_j)$, the linear constraints in (6) can be expressed as $\mathbf{A}\mathbf{m} = \mathbf{e}_1$. Then, the projection of \mathbf{M} onto \mathcal{C}_{lin} is given by $\text{vec}^{-1}(\mathbf{m} + \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}(\mathbf{e}_1 - \mathbf{A}\mathbf{m}))$. This expression simplifies significantly by close inspection. In fact, it is equivalent to setting $m_{11} = 1$ and for $j \geq 2$ replacing m_{j1} , m_{1j} , and m_{jj} by their average $(m_{jj} + m_{j1} + m_{1j})/3$. Denote this projection with the operator $\mathcal{P}_{lin}(\cdot)$.

Second, Tanaka and Nakata (2014) describe a univariate optimization problem that is almost equivalent to projecting \mathbf{M} onto \mathcal{C}_{sd} . The only difference is that their solution set also includes the zero-matrix $\mathbf{0}$. Assuming that such a limiting case can be safely handled in the implementation, their approach offers a fast projection onto \mathcal{C}_{sd} even for a moderately large N . To describe this approach, consider the spectral decomposition $\mathbf{M} = \mathbf{Q}\text{Diag}(l_1, \dots, l_{N+1})\mathbf{Q}'$

and the univariate function

$$\pi(\mu) = \sum_{i=1}^{N+1} [(\mu - l_i)_+^2 + (l_i - \kappa\mu)_+^2],$$

where $\text{Diag}(\mathbf{x})$ is a diagonal matrix with diagonal \mathbf{x} and $(\cdot)_+$ is the positive part operator. The function $\pi(\mu)$ can be minimized very efficiently by solving a series of smaller convex problems, each with a closed form solution. The result is a binary-search-like procedure described by Algorithm ?? in Appendix A. If $\mu^* = \arg \min_{\mu \geq 0} \pi(\mu)$ and

$$\lambda_j^* := \begin{cases} \mu^* & \text{if } l_j \leq \mu^* \\ \kappa\mu^* & \text{if } \kappa\mu^* \leq l_j \\ l_j & \text{otherwise} \end{cases}$$

for $j = 1, \dots, N + 1$, then $\mathbf{Q}\text{Diag}(\lambda_1^*, \dots, \lambda_{N+1}^*)\mathbf{Q}$ is the projection of \mathbf{M} onto \mathcal{C}_{sd} . Call this projection $\mathcal{P}_{sd}(\cdot : \kappa)$.

Algorithm 1 uses these projections to solve (6). Each iteration projects twice on one set and once on the other set. The general form of the algorithm does not specify which projection should be called twice. Therefore, given that $\mathcal{P}_{sd}(\cdot : \kappa)$ takes longer to run than $\mathcal{P}_{lin}(\cdot)$, it is beneficial to choose to call $\mathcal{P}_{lin}(\cdot)$ twice. The complexity of each iteration is determined largely by the spectral decomposition which is fairly fast for moderately large N . Overall time to convergence, of course, depends on the choice of the stopping criterion. Many intuitive criteria are possible. Given that $\mathbf{\Omega}_D \in \mathcal{C}_{lin}$ and $\mathbf{\Omega}_C \in \mathcal{C}_{sd}$, the stopping criterion $\max\{(\mathbf{\Omega}_D - \mathbf{\Omega}_C)_{ij}^2\} < \epsilon$ suggests that the return value is in \mathcal{C}_{sd} and close to \mathcal{C}_{lin} in every direction. Based on our experience, the algorithm converges quite quickly. For instance, our implementation in C++ generally solves (6) for $\epsilon = 10^{-5}$ and $N = 100$ in less than a second on a 1.7 GHz Intel Core i5 computer. This code will be made available online upon publication. For the remainder of

Require: Unconstrained covariance matrix estimator \mathbf{S} , stopping criterion $\epsilon > 0$, and an upper bound on the condition number $\kappa \geq 1$.

```

1: procedure DIRECTIONAL ALTERNATING PROJECTION ALGORITHM
2:    $\Omega_A \leftarrow h(\mathbf{S})$ 
3:   repeat
4:      $\Omega_B \leftarrow \mathcal{P}_{lin}(\Omega_A)$ 
5:      $\Omega_C \leftarrow \mathcal{P}_{sd}(\Omega_B : \kappa)$ 
6:      $\Omega_D \leftarrow \mathcal{P}_{lin}(\Omega_C)$ 
7:      $\Delta \leftarrow \|\Omega_B - \Omega_C\|_F^2 / \text{tr}[(\Omega_B - \Omega_D)'(\Omega_B - \Omega_C)]$ 
8:      $\Omega_A \leftarrow \Omega_B + \Delta(\Omega_D - \Omega_B)$ 
9:   until  $\max \left\{ (\Omega_D - \Omega_C)_{ij}^2 \right\} < \epsilon$ 
10:  return  $h^{-1}(\Omega_C)$ 
11: end procedure

```

Algorithm 1: This procedure projects $h(\mathbf{S})$ onto the intersection $\mathcal{C}_{sd} \cap \mathcal{C}_{lin}$. Denote the projection with $\mathcal{P}_{LSE}(\mathbf{S} : \kappa)$. Throughout the paper, the stopping criterion is fixed at $\epsilon = 10^{-5}$.

the paper, projecting \mathbf{S} onto the feasible region is denoted with the operator $\mathcal{P}_{LSE}(\mathbf{S} : \kappa)$.

3.4 Selecting κ

The estimation procedure described in the previous section has one tuning parameter, namely the condition number threshold κ . This subsection discusses an in-sample approach, called *conditional validation*, that can be used for choosing any tuning parameter, such as κ , under the partial information framework. To motivate, recall that the revealed aggregator X'' uses Σ to regress Z_0 on the rest of the Z_j s. Of course, the accuracy of this prediction cannot be known until the actual outcome is observed. However, apart from being unobserved, the variable Z_0 is theoretically no different to the other Z_j s. This suggests the following algorithm: for some value ν compute $\mathcal{P}_{LSE}(\mathbf{S} : \nu)$, let each of the Z_j s in turn play the role of Z_0 , predict its value based on Z_i for $i \neq j$, and choose the value of ν that yields the best overall accuracy. Even though many accuracy measures could be chosen, this paper uses the conditional log-likelihood. Therefore, if $\mathbf{Z}_j^* = (Z_{j1}, \dots, Z_{jK})'$ collects the j th forecaster's information about

the K events, the chosen value of κ is

$$\kappa_{cov} = \arg \max_{\nu \geq 1} \sum_{j=1}^N \ell(\mathbf{Z}_j^*, \mathcal{P}_{LSE}(\mathbf{S} : \nu) | \mathbf{Z}_i^* \text{ for } i \neq j), \quad (7)$$

where the log-likelihood is now conditional on \mathbf{Z}_i^* s for $i \neq j$ and \mathbf{S} is computed based on all the forecasts $\mathbf{Z}_1^*, \dots, \mathbf{Z}_N^*$. Plugging this into the projection algorithm gives the final estimate $\Sigma_{cov} := \mathcal{P}_{LSE}(\mathbf{S} : \kappa_{cov})$.

Unfortunately, the optimization problem (7) is non-convex in ν . However, as was mentioned before, Algorithm 1 is fast for moderately sized N . Therefore κ can be chosen efficiently (possibly in parallel on multicore machines) over a grid of candidate values. Overall, the idea in conditional validation is similar to cross-validation but, instead of predicting across rows (observations), the prediction is performed across columns (variables). This not only mimics the actual process of revealed aggregation but is also likely to be more appropriate for prediction polling that typically involves a large number of forecasters (large N) predicting relatively few events (small K). Furthermore, it has no tuning parameters and remains more stable when K is small; see Appendix B for an illustration of this result under synthetic data.

4. APPLICATIONS

This section applies the partial information framework to different types of real world forecasts. For each type there may be different ways to adopt the Gaussian model. The main point, however, is not to find the optimal way to do this but rather to give illustrative examples on using the framework and also to show how the resulting partial information aggregators outperform the commonly used measurement error aggregators.

4.1 Probability Forecasts of Binary Outcomes

4.1.1 Dataset

During the second year of the Good Judgment Project (GJP) the forecasters made probability estimates for 78 events, each with two possible outcomes. One of these events was illustrated in Figure 1. Each prediction problem had a timeframe, defined as the number of days between the first day of forecasting and the anticipated resolution day. These timeframes varied largely among problems, ranging from 12 days to 519 days with a mean of 185.4 days. During each timeframe the forecasters were allowed to update their predictions as frequently as they liked. The forecasters knew that their estimates would be assessed for accuracy using the quadratic loss (often known as the Brier score; see Brier 1950 for more details). This is a proper loss function that incentivized the forecasters to report their true beliefs instead of attempting to game the system. In addition to receiving \$150 for meeting minimum participation requirements that did not depend on prediction accuracy, the forecasters received status rewards for their performance via leader-boards displaying the losses for the best 20 forecasters. Depending on the details of the reward structure, such a competition for rank may eliminate the truth-revelation property of proper loss functions (see, e.g., Lichtendahl Jr and Winkler 2007).

This data collection raises several issues. First, given that the current paper does not focus on modeling dynamic data, only forecasts made within some common time interval should be considered. Second, not all forecasters made predictions for all the events. Furthermore, the forecasters generally updated their forecasts infrequently, resulting into a very sparse dataset. Such high sparsity can cause problems in computing the initial unconstrained estimator S . Evaluating different techniques to handle missing values, however, is well outside the scope of this paper. Therefore, to somewhat alleviate the effect of missing values, only the hundred most active forecasters are considered. This makes sufficient overlap highly likely but, unfortunately, still not guaranteed.

All these considerations lead to a parallel analysis of three scenarios: High Uncertainty (HU), Medium Uncertainty (MU), and Low Uncertainty (LU). Important differences are summarized in Table 1. Each scenario considers the forecasters' most recent prediction within a different time interval. For instance, LU only includes each forecaster's most recent forecast during 30 – 60 days before the anticipated resolution day. The resulting dataset has 60 events of which 13 occurred. In the corresponding 60×100 table of forecasts, around 42 % of the values are missing. The other two scenarios are defined similarly.

Table 1: Summary of the three time intervals analyzed. Each scenario considers the forecasters' most recent forecasts within the given time interval. The value in the parentheses represent the number of events occurred. The final column shows the percent of missing forecasts.

Scenario	Time Interval	# of Events	Missing (%)
High Uncertainty (HU)	90 – 120	49 (10)	51
Medium Uncertainty (MU)	60 – 90	56 (14)	46
Low Uncertainty (LU)	30 – 60	60 (13)	42

4.1.2 Model Specification and Aggregation

The first step is to pick a link function and derive a Gaussian model for probability forecasts of binary events. Overall, this construction resembles in many ways the latent variable version of a standard probit model.

Model Instance. Identify the k th event with $Y_k \in \{0, 1\}$. These outcomes link to the information variables via the following function:

$$Y_k = g(Z_{0k}) = \begin{cases} 1 & \text{if } Z_{0k} > t_k \\ 0 & \text{otherwise,} \end{cases}$$

where $t_k \in \mathbb{R}$ is some threshold value. Therefore the link function $g(\cdot)$ is simply the indicator function $\mathbf{1}_{A_k}$ of the event $A_k = \{Z_{0k} > t_k\}$. This threshold is defined

by the prior probability of the k th event $\mathbb{P}(Y_k = 1) = \Phi(-t_k)$, where $\Phi(\cdot)$ is the CDF of a standard Gaussian distribution. Given that the thresholds are allowed to vary among the events, each event has its own prior. The corresponding probability forecasts $X_{jk} \in [0, 1]$ are

$$X_{jk} = \mathbb{E}(Y_k | Z_{jk}) = \Phi \left(\frac{Z_{jk} - t_k}{\sqrt{1 - \delta_j}} \right).$$

In a similar manner, the revealed aggregator $X_k'' \in [0, 1]$ for event k is

$$X_k'' = \mathbb{E}(Y_k | \mathbf{Z}_k) = \Phi \left(\frac{\text{diag}(\Sigma)' \Sigma^{-1} \mathbf{Z}_k - t_k}{\sqrt{1 - \text{diag}(\Sigma)' \Sigma^{-1} \text{diag}(\Sigma)}} \right). \quad (8)$$

All the parameters of this model can be estimated from the data. The first step is to specify a version of the unconstrained estimate \mathbf{S} . If the t_k 's do not change much, a reasonable and simple estimate is obtained by transforming the sample covariance matrix \mathbf{S}_P of the probit scores $P_{jk} := \Phi^{-1}(X_{jk})$. More specifically, if $\mathbf{D} := \text{Diag}(\mathbf{d})\text{Diag}(\mathbf{1} + \mathbf{d})^{-1}$, where $\mathbf{d} = \text{diag}(\mathbf{S}_P)$, then an unconstrained estimator of Σ is given by $\mathbf{S} = (\mathbf{I}_N - \mathbf{D})^{1/2} \mathbf{S}_P (\mathbf{I}_N - \mathbf{D})^{1/2}$. Recall that the GJP data holds many missing values. This is handled by estimating each pairwise covariance in \mathbf{S}_P based on all the events for which both forecasters made predictions. Next, compute Σ_{cov} , where κ_{cov} is chosen over a grid of 100 candidate values between 10 and 1,000. Finally, the threshold t_k can be estimated by letting $\mathbf{P}_k = (P_{1k}, \dots, P_{Nk})'$, observing that $-\text{Diag}(\mathbf{1} - \text{diag}(\Sigma))^{1/2} \mathbf{P}_k \sim \mathcal{N}(t_k \mathbf{1}_N, \Sigma)$, and computing the precision-weighted average:

$$\hat{t}_k = - \frac{\mathbf{P}_k' \text{Diag}(\mathbf{1} - \text{diag}(\Sigma_{cov}))^{1/2} \Sigma_{cov}^{-1} \mathbf{1}}{\mathbf{1}' \Sigma_{cov}^{-1} \mathbf{1}}.$$

If \mathbf{P}_k has missing values, the corresponding rows and columns of Σ_{cov} are dropped. Intuitively, this estimator gives more weight to the forecasters with very little information. These estimates

are then plugged in to (8) to get the revealed aggregator X''_{cov} .

This aggregator is benchmarked against the state-of-the-art measurement-error aggregators, namely the average probability, median probability, average probit-score, and average log-odds. Unequally weighted averages were not considered because it is unclear how the weights would be determined based on forecasts alone, and even if this could be done somehow (perhaps based on self-assessment or organizational status), using unequal weights often leads to no or very small performance gains (Rowse et al., 1974; Ashton and Ashton, 1985; Flores and White, 1989). To avoid infinite log-odds and probit scores, extreme forecasts $X_{jk} = 0$ and 1 were censored to $X_{jk} = 0.001$ and 0.999, respectively. The results remain insensitive to the exact choice of censoring as long as this is done in a reasonable manner to keep the extreme probabilities from becoming highly influential in the logit- or probit-space. The accuracy of the aggregates is measured with the average root-mean-squared-error (RMSE). Note that this is nothing but the square root of the commonly used Brier score. Instead of considering all the forecasts at once, the aggregators are evaluated under different N via repeated subsampling of the 100 most active forecasters; that is, choose N forecasters uniformly at random, aggregate their forecasts, and compute the RMSE. This is repeated 1,000 times with $N = 5, 10, \dots, 65$ forecasters. Due to high computational cost, the simulation was stopped after $N = 65$. In the rare occasion where no pairwise overlap is available between one or more pairs of the selected forecasters, the subsampling is repeated until all pairs have at least one problem in common.

Figure 3 shows the average RMSEs under the three scenarios described in Table 1. Here a reasonable upper bound is given by 0.5 as this is the RMSE one would receive by constantly predicting 0.5. All presented scores, however, are well below it and improve uniformly from left to right, that is, from HU to LU. This reflects the decreasing level of uncertainty. In all the figures the measurement-error aggregators rank in the typical order (from worst to best): average probability, median probability, average probit, and average log-odds. Regardless of the level of uncertainty, the revealed aggregator X''_{cov} outperforms the averaging aggregators as

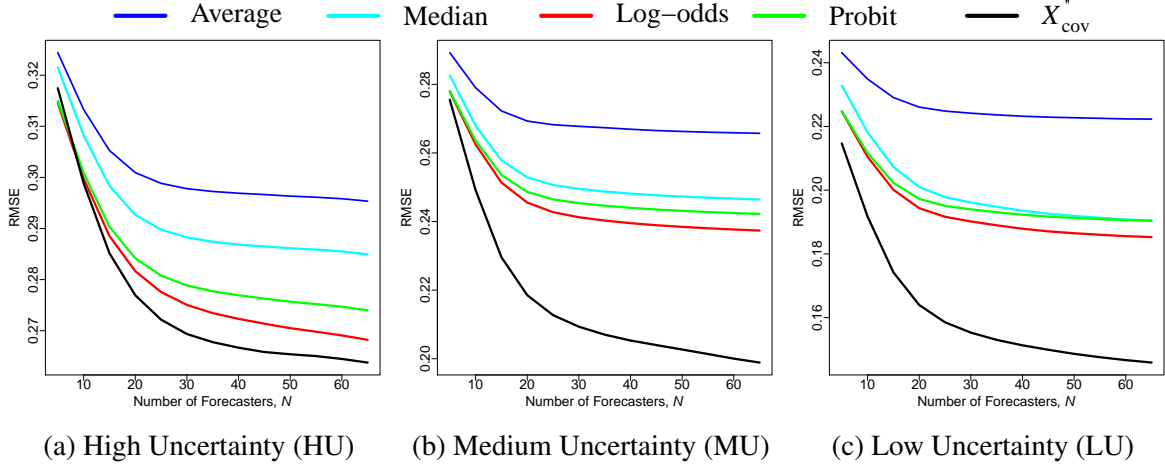


Figure 3: Average prediction accuracy over the 1,000 sub-samplings of the forecasters. See Table 1 for descriptions of the different scenarios.

long as $K \geq 10$. The relative advantage, however, increases from HU to LU. More specifically, the improvement from Log-odds to X''_{cov} is about 2%, 17%, and 21% in HU, MU, and LU, respectively. This trend can be explained by several reasons. First, as can be seen in Table 1, the amount of data increases from HU to LU. This yields a better estimate of Σ and hence more accurate revealed aggregation. Second, the forecasters are more likely to be well-calibrated under MU and LU than under HU (see, e.g., Braun and Yaniv 1992). Third, under HU the events are still inherently very uncertain. Consequently, the forecasters are unlikely to hold much useful information as a group. Under such low information diversity, measurement-error aggregators generally perform relatively well (Satopää et al. 2015). In the contrary, under MU the events have lost a part of their inherent uncertainty, allowing some forecasters to possess useful private information. These individuals are then prioritized by X''_{cov} while the averaging-aggregators continue treating all forecasts equally. Consequently, the performance of the measurement error aggregators plateaus after $N = 30$ or so. Therefore having more than about 30 forecasters does not make a difference if one is determined to aggregate their predictions using the measurement error techniques; a similar results was reported by Satopää et al. 2014a. In

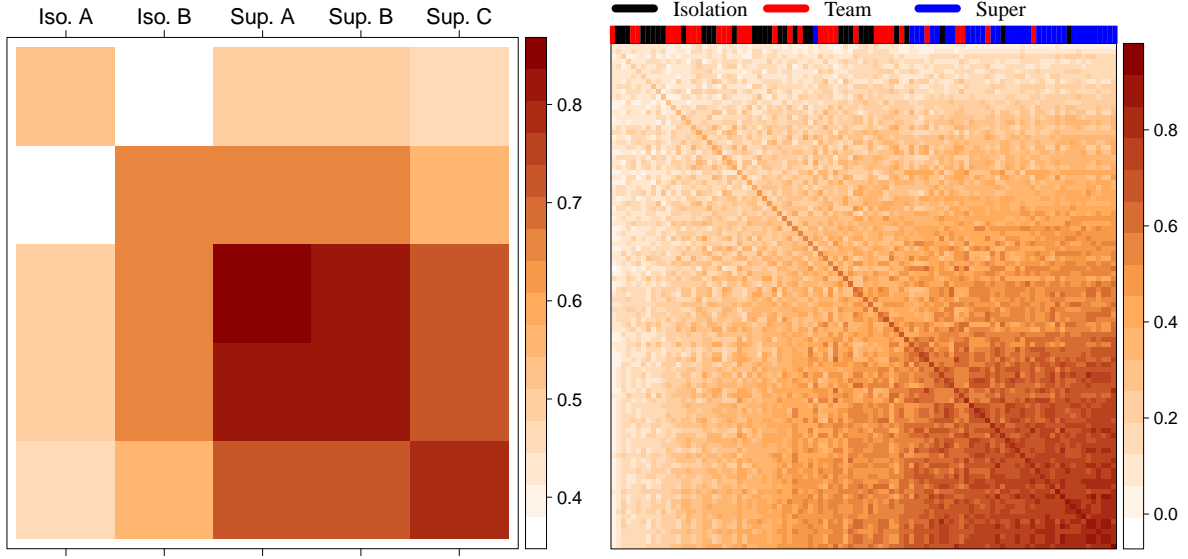
contrast, however, the RMSE of X''_{cov} continues to improve linearly in N , suggesting that X''_{cov} is able to find some residual information in each additional forecaster and use this to increase its performance advantage.

4.1.3 Information Diversity

The GJP assigned the forecasters to make predictions either in isolation or in teams. Furthermore, after the first year of the tournament, the top 2% forecasters were elected to the elite group of “super-forecasters.” These super-forecasters then worked in exclusive teams to make highly accurate predictions on the same events as the rest of the forecasters. Overall, these assignments directly suggest a level of information overlap. In particular, recalling the interpretation of Σ from Section 2.2.1, super-forecasters can be expected to have the highest δ_j s and forecasters in the same team should have a relatively high ρ_{ij} . This subsection examines how well Σ_{cov} aligns with this prior knowledge about the forecasters’ information structure.

For the sake of brevity, only the LU scenario is analyzed as this is where X''_{cov} presented the highest relative improvement. The associated 100 forecasters involve 36 individuals predicting in isolation, 33 forecasting team-members (across 24 teams), and 31 super-forecasters (across 5 teams). Figure 4a displays Σ_{cov} for the five most active forecasters. This group involves two forecasters working in isolation (Iso. A and B) and three super-forecasters (Sup. A, B, and C), of whom the super-forecasters A and B are in the same team. Overall, Σ_{cov} agrees with this classification: the only two team members, namely Sup. A and B have a relatively high information overlap. In addition, the three super-forecasters are more informed than the non-super-forecasters. Such a high level of information unavoidably leads to higher information overlap with the rest of the forecasters.

By and large, this agreement generalizes to the entire group of forecasters. To illustrate, Figure 4b displays Σ_{cov} for all the 100 forecasters. The information structure has been ordered with respect to the diagonal such that the more informed forecasters appear on the right. Fur-



(a) Σ_{cov} for the five most active forecasters

(b) Σ_{cov} for all 100 forecasters shows high information diversity.

Figure 4: The estimated information structure Σ under the LU scenario. Each forecaster worked either in isolation, in a non-super-forecaster team, or in a super-forecaster team. The super-forecasters generally have more information than the forecasters working in isolation.

thermore, a colored rug has been appended on the top. This rug shows whether each forecaster worked in isolation, in a non-super-forecaster team, or in a super-forecaster team. Observe that the super-forecasters are mostly situated on the right among the most informed forecasters. The average estimated δ_j among the super-forecaster is 0.80. On the other hand, the average estimated δ_j among the individuals working in isolation or in non-super-forecaster teams are 0.47 and 0.50, respectively. Therefore working in a team makes the forecasters' predictions, on average, slightly more informed.

In general, a plot such as Figure 4b is useful for assessing the level of information diversity among the forecasters: the further away it is from a monochromatic plot, the higher the

information diversity. That being said, the colorful Figure 4b suggests that the GJP forecasters have high information diversity. This makes sense as these forecasters were asked to make predictions about international political events. Given that on such events the forecasters' background knowledge, education, how closely they follow the news, and so on matter, one should expect a high level of information diversity. Therefore not only does X''_{cov} clearly outperform the common measurement error aggregators in terms of prediction accuracy but the Gaussian model also captures true structure in the data.

4.2 Point Forecasts of Continuous Outcomes

4.2.1 Dataset

Moore and Klein (2008) hired 415 undergraduates from Carnegie Mellon University to guess the weights of 20 people based on a series of pictures. These forecasts were illustrated in Figure 2. The target people were between 7 and 62 years old and had weights ranging from 61 to 230 pounds, with a mean of 157.6 pounds. All the students were shown the same pictures and hence given the exact same information. Therefore any information diversity arises purely from the participants' decisions to use different subsets of the same information. Consequently, information diversity is likely to be low compared to Section 4.1 where diversity also stemmed from differences in the information available to the forecasters.

Unlike in Section 4.1, the Gaussian model can be applied almost directly to the data. Only the effect of extreme values was reduced via a 90% Winsorization (Hastings et al., 1947). This handled some obvious outliers. For instance, the original dataset contained a few estimates above 1000 pounds and as low as 10 pounds. Winsorization generally improved the performance of all the competing aggregators.

4.2.2 Model Specification and Aggregation

Model Instance. Suppose Y_k and X_{jk} are real-valued. If the proper non-informative prior distribution of Y_k is $\mathcal{N}(\mu_{0k}, \sigma_0^2)$, then $Y_k = g(Z_{0k}) = Z_{0k}\sigma_0 + \mu_{0k}$. Consequently, $X_{jk} = \mathbb{E}(Y|Z_{jk}) = Z_{jk}\sigma_0 + \mu_{0k}$ for all $j = 1, \dots, N$. Therefore $X_j \sim \mathcal{N}(\mu_{0k}, \sigma_j^2)$ for some $\sigma_j^2 \leq \sigma_0^2$. If $\mathbf{Z}_k = (Z_{1k}, \dots, Z_{Nk})'$, then the revealed aggregator for the k th event is

$$X_k'' = \mathbb{E}(Y_k|\mathbf{Z}_k) = \text{diag}(\Sigma)' \Sigma^{-1} \mathbf{Z}_k \sigma_0 + \mu_{0k}. \quad (9)$$

Under this model the prior distribution of Y_k is specified by μ_{0k} and σ_0^2 . Given that $\mathbb{E}(X_{jk}) = \mu_{0k}$ for all $j = 1, \dots, N$, the sample average $\hat{\mu}_{0k} = \sum_{j=1}^N X_{jk}/N$ provides an initial estimate of μ_{0k} . The value of σ_0^2 can be estimated by assuming a distribution for the σ_j^2 s. More specifically, let σ_j^2 be i.i.d. on the interval $[0, \sigma_0^2]$ and use the resulting likelihood to estimate σ_0^2 . For instance, a non-informative choice is to assume $\sigma_j^2 \stackrel{i.i.d.}{\sim} \mathcal{U}(0, \sigma_0^2)$, which leads to the maximum likelihood estimator $\max\{\sigma_j^2\}$. This has a downward bias that can be corrected by a multiplicative factor of $(N+1)/N$. Therefore, replacing σ_j^2 with the sample variance $s_j = \sum_{k=1}^K (X_{jk} - \hat{\mu}_{0k})^2 / (K-1)$ gives the final estimate $\hat{\sigma}_0^2 = (N+1)/N \max\{s_j\}$. Using these estimates, the X_{jk} s can be transformed into the Z_{jk} s whose sample covariance matrix \mathbf{S}_Z provides the unconstrained estimator for the projection algorithm. The value of κ_{cov} is chosen over a grid of 10 values between 10 and 10,000. Once Σ_{cov} has been computed, the prior means are updated with the precision-weighted averages $\hat{\mu}_{0k} = (\mathbf{X}'_k \Sigma_{cov}^{-1} \mathbf{1}_N) / (\mathbf{1}'_N \Sigma_{cov}^{-1} \mathbf{1}_N)$. In the end, all these estimates are plugged in (9) to get the revealed aggregator X_{cov}'' .

This aggregator is compared against the average, median, and average of the median and average (AMA). The last competitor, namely AMA is a heuristic aggregator that Lobo and Yao (2010) showed to work particularly well on many different real-world forecasting datasets. In this section the overall accuracy is measured with the RMSE averaged over 10,000 sub-

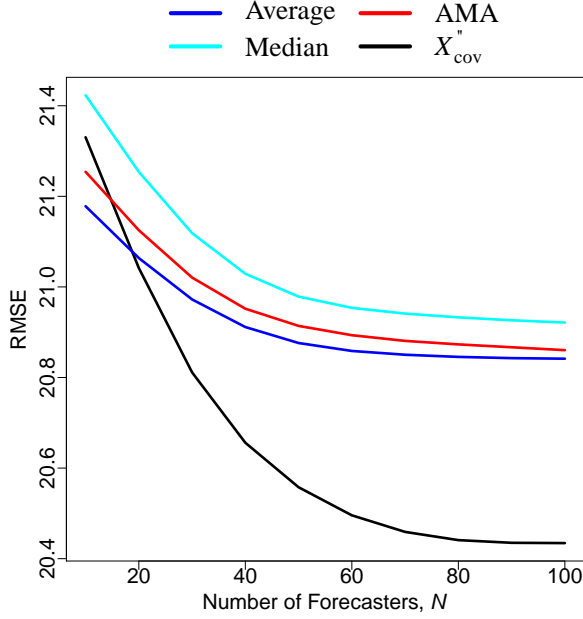


Figure 5: Average prediction accuracy

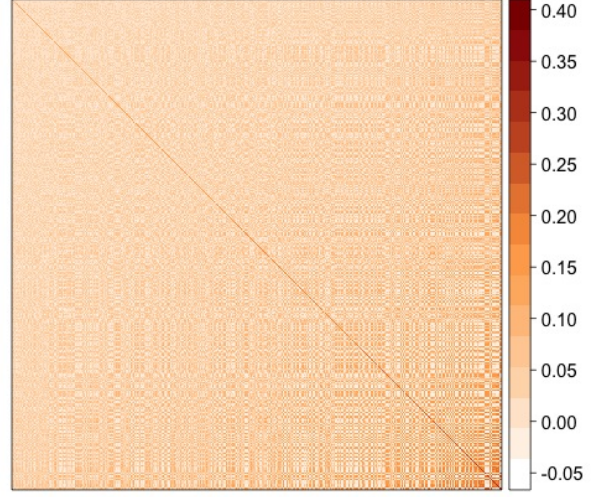


Figure 6: Σ_{cov} for all 416 forecasters shows low information diversity.

samplings of the 416 participants. That is, each iteration chooses N participants uniformly at random, aggregates their forecasts, and computes the RMSE. The size of the sub-samples is varied between 10 and 100 with increments of 10. These scores are presented in Figure 5. The average outperforms the median across all N . The performance of AMA falls between that of average and median, reflecting its nature as a compromise of the two. The revealed aggregator X''_{cov} is the most accurate once $N > 10$. The relatively worse performance at $N = 10$ suggests that 10 observations is not enough to estimate $\hat{\mu}_{0k}$ accurately. As N approaches 100, however, X''_{cov} collects information efficiently and increases the performance advantage against the other aggregators.

Figure 6 shows Σ_{cov} for all the 416 forecasters. Similarly to before, the matrix has been ordered such that the most knowledgeable forecasters are on the right. Overall, this plot is much more monochromatic than the one presented earlier in Figure 4b, suggesting that information

diversity among the 416 students is rather lower. This aligns with the expectations laid out earlier in Section 4.2.1. If there were no information diversity, i.e., all the forecasters used the same information, then averaging aggregators, such as the simple average, would perform very well (Satopää et al., 2015). Such a limiting case, however, is rarely encountered in practice. Often at least some information diversity is present. The results in the current section show that the revealed aggregator does not require extremely high information diversity in order to outperform the measurement-error aggregators.

5. DISCUSSION

This paper introduced the partial information framework for modeling forecasts from different types of prediction polls. Even though the framework can be used for theoretical analysis and studying information among groups of experts, the main focus was on model-based aggregation of forecasts. Such aggregators do not require a training set. Instead, they operate under a model of forecast heterogeneity and hence can be applied to forecasts alone. Under the partial information framework, all forecast heterogeneity stems from differences in the way the forecasters use information. Intuitively, this is more plausible at the micro-level than the historical measurement error. To facilitate practical applications, the partial information framework motivates and describes the forecasters' information with a patterned covariance matrix (Equation 1). A correctional procedure was proposed (Algorithm 1) as a general tool for estimating these information structures. This procedure inputs any covariance estimator and modifies it minimally such that the final output represents a physically feasible allocation of information. Even though the general partial information framework describes an optimal aggregator, it is generally too abstract to be directly applied in practice. As a solution, this paper discusses a close yet practical specification within the framework, known as the Gaussian model (Section 2.2.2). The Gaussian model permits a closed-form solution for the optimal aggregator and extends to different types of forecast-outcome pairs via a link function. These partial information

aggregators were evaluated against the common measurement error aggregators on two different real-world (Section 4) prediction polls. In each case the Gaussian model outperformed the typical measurement-error-based aggregators, suggesting that information diversity is more important for modeling forecast heterogeneity.

Generally speaking, partial information aggregation works well because it downweights pairs or sets of forecasters that share more information and upweights ones that have unique information (or choose to attend to unique information as is the case, e.g., in Section 4.2, where forecasters made judgments based on the same pictures). This is very different from measurement-error aggregators that assume all forecasters to have the same information and hence consider them equally important. While simple measurement-error techniques, such as the average or median, can work well when the forecasters truly operate on the same information set, in real-world prediction polls participants are more likely to have unequal skill and information sets. Therefore prioritizing is almost certainly called for. Of course, the more diverse these sets are, the better the partial information aggregators can be expected to perform relative to the measurement error aggregators. To illustrate this result, compare the relative performances in Section 4.1 (high information diversity) against those in Section 4.2 (low information diversity).

Overall, the partial information framework can be applied and extended in many different ways. For instance, in this paper the j th forecaster’s prediction was assumed to be the expectation of Y after observing some partial information \mathcal{F}_j . In some applications, however, other constructs, such as the conditional median or other quantiles, may be more appropriate. Such extensions can be handled by considering the distribution of $Y|\mathcal{F}_j$ and then equating the j th forecaster’s prediction to any desired functional of this distribution. This is particularly easy under the Gaussian model, where $Y|\mathcal{F}_j$ conveniently follows a Gaussian distribution.

In terms of future research, the partial information framework offers both theoretical and empirical directions. One theoretical avenue involves estimation of information overlap. In

some cases the higher order overlaps have been found to be irrelevant to aggregation. For instance, DeGroot and Mortera (1991) show that the pairwise conditional (on the truth) distributions of the forecasts are sufficient for computing the optimal weights of a weighted average. Theoretical results on the significance or insignificance of higher order overlaps under the partial information framework would be desirable. Given that the Gaussian model can only accommodate pairwise information overlap, such a result would reveal the need of a specification that is more complex than the Gaussian model.

A promising empirical direction is the Bayesian approach. These techniques are very natural for fitting hierarchical models such as the ones discussed in this paper. Furthermore, in many applications with small or moderately sized datasets, Bayesian methods have been found to be more stable than the likelihood-based alternatives. Therefore, given that the number of forecasts in a prediction poll is typically quite small, a Bayesian approach is likely to improve the quality of the final aggregate. This would involve developing a prior distribution for the information structure – a problem that seems interesting in itself. Overall, this avenue should certainly be pursued, and the results tested against other high performing aggregators.

REFERENCES

- Ashton, A. H. and Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, 31(12):1499–1508.
- Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a bregman predictor. *Information Theory, IEEE Transactions on*, 51(7):2664–2669.
- Braun, P. A. and Yaniv, I. (1992). A case study of expert judgment: Economists’ probabilities versus base-rate model forecasts. *Journal of Behavioral Decision Making*, 5(3):217–231.
- Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.

- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.
- Broecker, J. (2012). *Forecast verification: a practitioner's guide in atmospheric science*, chapter 7.2.2, pages 121–122. John Wiley & Sons, Chichester, UK, 2nd edition.
- Broomell, S. B. and Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, 74(3):531–553.
- Dawid, A., DeGroot, M., Mortera, J., Cooke, R., French, S., Genest, C., Schervish, M., Lindley, D., McConway, K., and Winkler, R. (1995). Coherent combination of experts' opinions. *TEST*, 4(2):263–313.
- Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.
- DeGroot, M. H. and Mortera, J. (1991). Optimal linear opinion pools. *Management Science*, 37(5):546–558.
- Di Bacco, M., Frederic, P., and Lad, F. (2003). Learning from the probability assertions of experts. Research Report. Available at: <http://www.math.canterbury.ac.nz/research/ucdms2003n6.pdf>.
- Dowie, J. (1976). On the efficiency and equity of betting markets. *Economica*, 43(170):139–150.
- Flores, B. E. and White, E. M. (1989). Subjective versus objective combining of forecasts: an experiment. *Journal of Forecasting*, 8(3):331–341.
- Foster, D. P. and Vohra, R. V. (1998). Asymptotic calibration. *Biometrika*, 85(2):379–390.

- Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991). Probabilistic mental models: a brunswikian theory of confidence. *Psychological Review*, 98(4):506.
- Goel, S., Reeves, D. M., Watts, D. J., and Pennock, D. M. (2010). Prediction without markets. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 357–366. ACM.
- Gubin, L., Polyak, B., and Raik, E. (1967). The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24.
- Hastings, C., Mosteller, F., Tukey, J. W., and Winsor, C. P. (1947). Low moments for small samples: A comparative study of order statistics. *The Annals of Mathematical Statistics*, 18(3):413–426.
- Hong, L. and Page, S. (2009). Interpreted and generated signals. *Journal of Economic Theory*, 144(5):2174–2196.
- Hwang, S.-G. (2004). Cauchy’s interlace theorem for eigenvalues of hermitian matrices. *American Mathematical Monthly*, 111:157–159.
- Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one’s general knowledge. *European Journal of Cognitive Psychology*, 5(1):55–71.
- Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, 39(1):98–114.
- Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3):217–273.
- Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6(2):107.

- Kruglanski, A. W. (1990). *Motivations for judging and knowing: Implications for causal attribution*, volume 2, pages 333–368. Guilford Press, New York, NY, US.
- Langford, E., Schwertman, N., and Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician*, 55(4):322–325.
- Lichtendahl Jr, K. C. and Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, 53(11):1745–1755.
- Lichtenstein, S. and Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational behavior and human performance*, 20(2):159–183.
- Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1977). *Calibration of probabilities: The state of the art*, volume 16 of *Theory and Decision Library*, pages 275–324. Springer Netherlands.
- Lobo, M. S. and Yao, D. (2010). Human judgement is heavy tailed: Empirical evidence and implications for the aggregation of estimates and forecasts. Available at http://sousalobo.com/researchfiles/Lobo_Yao_MS_11.pdf. (Working paper).
- McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press, 2nd edition.
- Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., and Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115.
- Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502.

- Moore, D. A. and Klein, W. M. (2008). Use of absolute and comparative performance feedback in absolute and comparative judgments and decisions. *Organizational Behavior and Human Decision Processes*, 107(1):60–74.
- Murphy, A. H. and Daan, H. (1984). Impacts of feedback and experience on the quality of subjective probability forecasts. comparison of results from the first and second years of the zierikzee experiment. *Monthly Weather Review*, 112(3):413–423.
- Murphy, A. H. and Winkler, R. L. (1977a). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature. *National Weather Digest*, 2(2):2–9.
- Murphy, A. H. and Winkler, R. L. (1977b). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 26(1):41–47.
- Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338.
- Parunak, H. V. D., Brueckner, S. A., Hong, L., Page, S. E., and Rohwer, R. (2013). Characterizing and aggregating agent estimates. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, pages 1021–1028, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.
- Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.
- Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91.
- Ravishanker, N. and Dey, D. K. (2001). *A first course in linear model theory*. CRC Press.

- Rowse, G. L., Gustafson, D. H., and Ludke, R. L. (1974). Comparison of rules for aggregating subjective likelihood ratios. *Organizational Behavior and Human Performance*, 12(2):274–285.
- Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014a). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356.
- Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., Ungar, L. H., et al. (2014b). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *The Annals of Applied Statistics*, 8(2):1256–1280.
- Satopää, V. A., Pemantle, R., and Ungar, L. H. (2015). Modeling probability forecasts via information diversity. *The Journal of the American Statistical Association (Theory & Methods)* (*arXiv:1406.2148*) (*In Press*).
- Satopää, V. A. and Ungar, L. H. (2015). Combining and extremizing real-valued forecasts. *arXiv:1506.06405* (*Under Review*).
- Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.
- Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, 65(2):117–137.
- Tanaka, M. and Nakata, K. (2014). Positive definite matrix approximation with condition number constraint. *Optimization Letters*, 8(3):939–947.
- Ungar, L., Mellers, B., Satopää, V., Tetlock, P., and Baron, J. (2012). The good judgment

project: A large scale test of different methods of combining expert predictions. The Association for the Advancement of Artificial Intelligence Technical Report FS-12-06.

Won, J. H. and Kim, S.-J. (2006). Maximum likelihood covariance estimation with a condition number constraint. In *Signals, Systems and Computers, 2006. ACSSC'06. Fortieth Asilomar Conference on*, pages 1445–1449. IEEE.

Yates, J. F. (1990). *Judgment and decision making*. Prentice-Hall, Inc, illustrated edition.